

# Object-based World Modeling in Semi-Static Environments with Dependent Dirichlet-Process Mixtures

Lawson L.S. Wong, Thanard Kurutach, Leslie Pack Kaelbling, Tomás Lozano-Pérez  
CSAIL, MIT, Cambridge, MA 02139  
{ lsw, kurutach, lpk, tlp }@csail.mit.edu

## Abstract

To accomplish tasks in human-centric indoor environments, robots need to represent and understand the world in terms of objects and their attributes. We refer to this attribute-based representation as a world model, and consider how to acquire it via noisy perception and maintain it over time, as objects are added, changed, and removed in the world. Previous work has framed this as multiple-target tracking problem, where objects are potentially in motion at all times. Although this approach is general, it is computationally expensive. We argue that such generality is not needed in typical world modeling tasks, where objects only change state occasionally. More efficient approaches are enabled by restricting ourselves to such semi-static environments.

We consider a previously-proposed clustering-based world modeling approach that assumed static environments, and extend it to semi-static domains by applying a dependent Dirichlet-process (DDP) mixture model. We derive a novel MAP inference algorithm under this model, subject to data association constraints. We demonstrate our approach improves computational performance in semi-static environments.

## 1 Introduction

There are many situations in which it is important for an automated system to maintain an estimate of the state of a complex dynamical system. Many physical systems are well described in terms of a set of objects, attributes of those objects, and relations between them. The number and properties of the objects in the world may change over time, and they are only partially observable due to noise and occlusion in the observation process. Domains that are appropriately modeled this way include: a household robot, which must maintain an estimate of the contents of a refrigerator that is used by multiple other people based on partial views of its contents; a wildlife-monitoring drone, which must maintain an estimate of the number, age, and health of elephants in a herd based on a sequence of photos of the herd moving through a forest; a surveillance satellite, which must estimate the number, activity, and hostility of soldiers in an enemy camp based on photos capturing people only when they are outside of buildings.

Estimating properties of individuals from noisy observations is a relatively simple statistical estimation problem if the observations are labeled according to which individual generated them. Even when the underlying attributes of the individual change over time, the problem of inferring the history of each individual’s attributes can be reduced to a problem of inference in a hidden Markov model.

The key difficulty in the problems described above is *data association*. We do not know which particular individual is responsible for each observation and so determining an appropriate association of observations to individuals is key. The only information we have to help make such associations are noisy and partial observations, which may contain errors both in attribute values and in number.

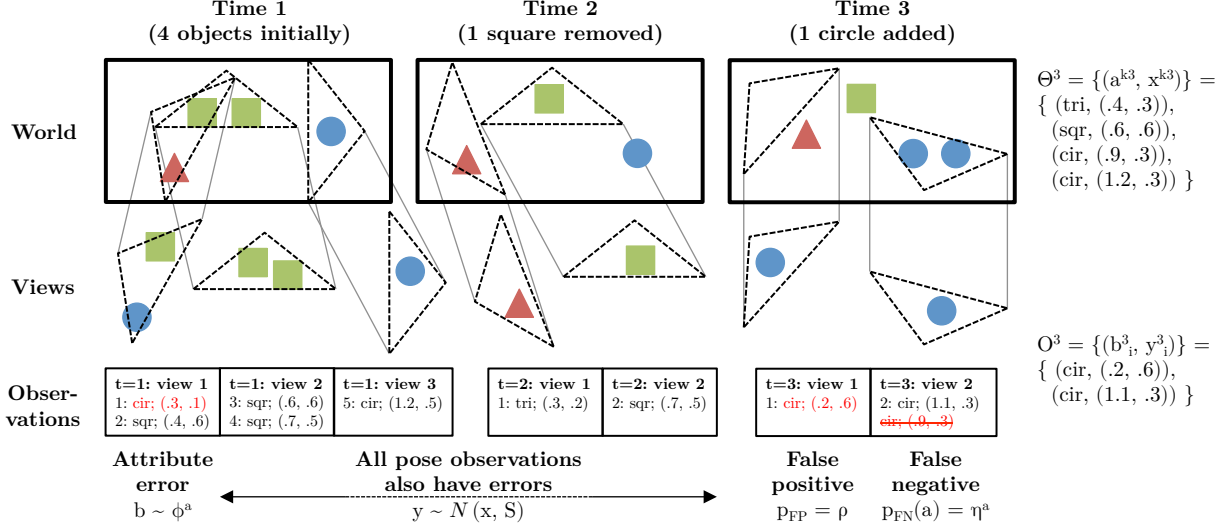
Within the context of world modeling, Cox and Leonard (1994) first identified this issue, and applied well-known multiple-hypothesis tracking (MHT) methods to resolve the issue (Reid, 1979; Bar-Shalom and Fortmann, 1988; Elfring et al., 2013). Recently, Oh et al. (2009) have pointed out drawbacks in using the MHT, which include inefficiency due to considering an exponential number of hypotheses, and the inability to revisit associations from previously-considered views (the MHT is essentially a forward filtering algorithm). Inspired by this, they and others (Dellaert et al., 2003; Pasula et al., 1999) have proposed different Markov-chain Monte Carlo (MCMC) methods for data association. See Wong et al. (2015) for in-depth coverage about previous work in semantic world modeling and data association.

The methods mentioned above were all formulated for multiple-target tracking problems, where the each target’s state (typically location) changes between observations. However, if we consider applications such as tracking objects in a household, the dynamics are typically different: most objects tend to stay in the same state when they are not being actively used. In this paper, we study the world modeling problem in *semi-static environments*, where time is divided into known epochs, and within each epoch the world is stationary. It seems intuitive that data association should be easier within static periods, since there is no uncertainty arising from stochastic dynamics.

An alternative approach to data association is to perform inference over the entire time-series of observations and to think of it as a problem of clustering: we wish to group together similar detections over time, under the assumption that they will have been generated by the same individual. Bayesian nonparametric models, such as the Dirichlet-process mixture model (DPMM), can be used to model domains in which the number of individuals is unknown *a priori*; in Wong et al. (2015), we found that a state-estimation technique based on DPMM clustering was effective for determining the number and type of objects in a static domain, given a sequence of images with partial views of the scene and significant occlusion.

In this paper, we apply the clustering approach to the much more difficult case of a dynamic domain in which the attributes of objects may change over time, new objects may appear, and old objects may permanently disappear. The DPMM is not an appropriate model for this problem, but an extension, the *dependent Dirichlet process mixture model* (DDPMM), which models dependencies between a collection of clusters, can be used effectively. In particular, we use a construction proposed by Lin et al. (2010) for a class of DDPs that can be represented as a Markov chain over DPs. In our case of semi-static world modeling, we model objects in each static epoch as clusters in a DPMM, and clusters between epochs are related by Markovian transitions, thus forming a DDPMM.

In the remainder, we will formalize the world modeling problem, review the DDP construction and apply it to our problem, and derive a novel maximum *a posteriori* (MAP) inference algorithm for the model. We show that this model yields computational advantages for tracking in semi-static environments, both in simulation and on real-world data.



**Figure 1:** An illustration of the world modeling problem. An unknown number of objects exist in the world (top row), and change in pose and number over time (world at each epoch enclosed in box). At each epoch, limited views of the world are captured, as depicted by the triangular viewcones. Within these viewcones, objects and their attributes are detected using black-box perception modules (e.g., off-the-shelf object detectors). In this example, the attributes are shape type (discrete) and 2-D location. The observations are noisy, as depicted by the perturbed versions of viewcones in the middle row. Uncertainty exists both in the attribute values and the existence of objects, as detections may include false positives and negatives (e.g.,  $t = 3$ ). The actual attribute detection values obtained from the views are shown in the bottom row (“Observations”); this is the format of input data. Given these noisy measurements as input, the goal is to determine which objects were in existence at each epoch, their attribute values (e.g.,  $\Theta^3$  in top right), and their progression over time.

## 2 Problem Definition

In world modeling, we seek the state of the world, consisting an unknown finite number  $K^t$  of objects, which changes over time. Object  $k$  at epoch  $t$  has attribute values  $\theta^{kt}$ . We sometimes decompose  $\theta^{kt}$  into  $(a^k, x^{kt})$ , where  $a$  is a vector of fixed attributes, and  $x$  is a vector of attributes that may change between epochs. The top row in Figure 1 illustrates the world state over three epochs for a simple domain.

Our system obtains noisy, partial views of the world. Each view  $v$  produces a set of observations  $O^{tv} = \{o_i^{tv}\}$ , where  $o_i^{tv} = (b_i^{tv}, y_i^{tv})$ , corresponding to the fixed attributes  $a$  and dynamic attributes  $x^t$  of some (possibly non-existent) object<sup>1</sup>. Each view is also associated with a field of view  $V^{tv}$ . The collection of views in a single epoch may fail to cover the entire world. The partial views and noisy observations are illustrated in the middle and bottom rows of Figure 1.

The world modeling problem can now be defined: Given observations  $O = \{o_i^{tv}\}_{(t,v,i)}$  and fields of view  $\{V^{tv}\}_{(t,v)}$ , determine the state of objects over time  $\Theta = \{\theta^{kt}\}_{(k,t)}$ . The state includes not

<sup>1</sup> Superscripts in variables will generally refer to the ‘context’, such as object index  $k$  and time index  $t$ . Subscripts refer to the index in a list, such as  $o_i^t = i$ ’th observation at time  $t$ .

only objects’ attribute values, but also the total number of objects that existed at each epoch, and implicitly when objects were added and removed (if at all).

There is no definitive information in the observations that will allow us to know which particular observations correspond with which underlying objects in the world, or even how many objects were in existence at any time step. For example, in the views of  $t = 1$  shown in Figure 1, the square detected in the left-most view may correspond to either (or neither) square in the center view. Also, despite there being only four objects in the world, there were five observations because of overlapping visible regions.

The critical piece of information that is missing is the *association*  $z_i^{tv}$  of an observation  $o_i^{tv}$  to an underlying object  $k$ . With this information, we can perform statistical aggregation of the observations assigned to the same object to recover its state. We will model the associations  $Z = \{z_i^{tv}\}_{(t,v,i)}$  as latent variables in a Bayesian inference process.

## 2.1 Observation noise model

The observation model describes how likely an observation  $o = (b, y)$  was generated from some given object state  $\theta = (a, x)$  (if any), given by the probability  $f(o; \theta)$ . For a single object, let  $\theta_c$  and  $\theta_d$  be the true continuous and discrete attribute values respectively, and likewise  $o_c$  and  $o_d$  for a single observation of the object. We typically consider observation noise models of the following form:

$$f(o; \theta) = \phi^{\theta_d}(o_d) \mathcal{N}(o_c; \theta_c, S) \quad (1)$$

Here  $\phi$  represents a discrete confusion matrix, where  $\phi^{\theta_d}(o_d)$  is the probability of observing  $o_d$  given the true object has discrete attributes  $\theta_d$ . The continuous-valued observation  $o_c$  is the true value  $\theta_c$  corrupted with zero-mean Gaussian noise, with fixed sensing covariance  $S$ . The noise on  $o_c$  and  $o_d$  are assumed to be independent for simplicity.

Besides errors in attribute values, Figure 1 also illustrates cases of false positives and false negatives. A false positive occurs when the observation did not originate from any true object. We assume that this occurs at a fixed rate  $\rho$ , depending on the perception system. When this occurs,  $o_d$  has noise distribution  $\phi^0$ , and  $o_c$  is uniformly distributed over the field of view  $V$ . A false negative occurs when an object is within the sensor’s field of view but failed to be detected. We assume that an object within the field of view  $V$  will be undetected with an attribute-dependent probability  $\eta(\theta)$ .

## 2.2 Additional assumption: Cannot-link constraint (CLC)

Finally, there is an additional common domain assumption in target-tracking problems that is essential: within a single view, each visible object can generate at most one detection Bar-Shalom and Fortmann (1988). This implies that within a view, each observation must be assigned to a different hypothesized underlying object. Adopting the terminology of clustering, we refer to this as a “cannot-link constraint” (CLC). The constraint is powerful because it can reduce ambiguities when there are similar nearby objects. However, clustering algorithms typically cannot handle such constraints, and similar to the DPMM-based data association work of Wong et al. (2015), we will need to modify the DDP model and inference algorithms to handle the world modeling problem.

### 3 A Clustering-Based Approach

We now specify a prior on how likely an assignment to a cluster is, and how clusters change over time. Since the number of clusters are unknown, we chose to use Bayesian nonparametric mixture models, which allow for an indefinite and unbounded number of mixture components. (although the number of instantiated components is limited by the data size).

The Dirichlet process (DP) (Teh (2010) provides a good overview), and its application to mixture modeling (Antoniak, 1974; Neal, 2000), is a widely-studied prior for density estimation and clustering. The DP’s popularity stems from its simplicity and elegance. However, one major limitation is that clusters cannot change over time, a consequence of the fact that observations are assumed to be fully exchangeable. This assumption is violated for problems like ours, where the observed entities change over time and space. Indeed, the previous application of DPs to world modeling mentioned above required that the world is static, which is a significant limitation. Various generalizations of the DP that model temporal dynamics have thus been proposed (Zhu et al., 2005; Ahmed and Xing, 2008).

Many of these generalizations belong to a broad class of stochastic models known as dependent Dirichlet processes (DDP) (MacEachern, 1999, 2000). We will adopt a theoretically-appealing instance of the DDP, based on a recently-proposed Poisson-process construction (Lin et al., 2010; Lin, 2012). This construction subsumes a number of existing algorithmically-motivated DP generalizations. Additionally, Lin’s construction has the nice property that at each time slice, the prior over clusters is marginally a DP. Given a DP prior at time  $t$ , the construction specifies a dependent prior at time  $t + 1$  (or another future time), which is shown to also be a DP. The construction therefore generates a Markov chain of DPs over time, which reflects temporal dynamics between epochs in our problem.

We now state one result of the DDP construction; see Appendix A, and Lin (2012) for details. The construction results in the following prior on parameter  $\theta^t$  (to be assigned to a new observation), given past parameters  $\Theta^{<t}$  and parameters  $\Theta^t$  corresponding to clusters that have already been instantiated at the current epoch:

$$\begin{aligned} \theta^0 \mid \Theta^0 &\propto \alpha H(\theta^0) + \sum_k N^{k0} \mathbb{I}[\theta^0 = \theta^{k0}] \\ \theta^t \mid \Theta^{\leq t} &\propto \alpha H(\theta^t) + \sum_{k: N^{kt} > 0} N^{k, \leq t} \mathbb{I}[\theta^t = \theta^{kt}] + \sum_{k: N^{kt} = 0} q(\theta^{k, t-1}) N^{k, < t} T(\theta^t; \theta^{k, t-1}) \end{aligned} \quad (2)$$

At the initial time step, clusters are formed as in a standard DPMM with concentration parameter  $\alpha$  and base distribution  $H$ . For later time steps, the prior distribution on  $\theta$  is defined recursively. The first two terms are similar to the base case, for new clusters and already-instantiated clusters (in the current epoch) respectively. The third term corresponds to previously-existing clusters that may be removed with probability  $(1 - q(\theta^{k, t-1}))$ , and, if it survives, is moved with transition probability  $T(\cdot; \theta^{k, t-1})$ .  $N^{k, \leq t}$  is the number of points that have been assigned to cluster  $k$ , for all time steps up to time  $t$ . This term is similar to that in the DP. Note that if  $q \equiv 1$  and  $T(\cdot; \theta) = \delta_\theta$ , then the model is static, and Equation 2 is equivalent to the predictive distribution in the DP.

#### 3.1 Inference by forward sampling

As mentioned in the problem definition, our focus will be on determining latent assignments  $Z = \{z_i^t\}$  of observations  $O = \{o_i^t\}$  to clusters with parameters  $\Theta = \{\theta^{kt}\}$ . In the generic DDP, views

do not exist yet; those will be introduced in Section 4. One way to explore the distribution of assignments is to sample repeatedly from the assignment's conditional distribution, given all other assignments  $Z_{\setminus ti} \triangleq Z \setminus \{z_i^t\}$ :

$$\begin{aligned} \mathbb{P}(z_i^t = k \mid o_i^t, \Theta, Z_{\setminus ti}) &= \int \mathbb{P}(z_i^t = k, \theta \mid o_i^t, \Theta, Z_{\setminus ti}) \, d\theta \\ &\propto \int \mathbb{P}(o_i^t \mid \theta) \mathbb{P}(\theta = \theta^{kt} \mid \Theta, Z_{\setminus ti}) \, d\theta \end{aligned} \quad (3)$$

The first term in the integrand is given by the observation noise model (Equation 1), and the second term is given by the DDP prior (Equation 2). If  $\theta^{kt}$  already exists, then  $\mathbb{P}(\theta \mid \Theta, Z_{\setminus ti}) = \mathbb{I}[\theta = \theta^{kt}]$ , and the integrand only has support for  $\theta = \theta^{kt}$ . Otherwise, we have to consider all possible settings of  $\theta^{kt}$ , which has a prior distribution given by Equation 2. The expression in Equation 3 above can be decomposed into three cases, corresponding to terms in Equation 2:

$$\mathbb{P}(z_i^t = k \mid o_i^t, \Theta^{\leq t}, Z_{\setminus ti}^{\leq t}) \propto \begin{cases} N_{\setminus ti}^{k, \leq t} & f(o_i^t; \theta^{kt}) , \\ & k \text{ existing, instantiated at } t \\ \tilde{q}(\theta^{k\tau}) N_{\setminus ti}^{k, < t} & \int f(o_i^t; \theta) \tilde{T}(\theta; \theta^{k\tau}) \, d\theta , \\ & k \text{ existing, not instantiated at } t \\ \alpha & \int f(o_i^t; \theta) H(\theta) \, d\theta , \\ & k \text{ new} \end{cases} \quad (4)$$

In the DDPMM, clusters move around the parameter space during their lifetimes, and, depending on our chosen viewpoints, may not generate observations at some epochs. When cluster  $k$  has at least one time- $t$  observation assigned to it, it becomes *instantiated* at time  $t$ . Any subsequent observations at time  $t$  that are assigned to cluster  $k$  must then share the same parameter  $\theta^{kt}$ ; this corresponds to the first case. The second case is for clusters not yet instantiated at time  $t$ , and we must infer  $\theta^{kt}$  from the last known parameter for cluster  $k$ , at time  $\tau < t$ . If  $t - \tau > 1$ , we use generalized survival and transition expressions for our application:

$$\begin{aligned} \tilde{q}(\theta^{k\tau}) &\triangleq [q(\theta^{k\tau})]^{t-\tau} \\ \tilde{T}(\theta^{kt}; \theta^{k\tau}) &= \mathbb{I}[a^{kt} = a^{k\tau}] \mathcal{N}(x^{kt}; x^{k\tau}, (t - \tau)R(a^k)) \end{aligned} \quad (5)$$

The third case is for new clusters that are added at time  $t$ . The first and third cases essentially have the same form as the Gibbs sampler for the (static) DP.

In general, since the cluster parameters  $\Theta$  are also unknown, inference schemes need to alternate between sampling the cluster assignments (given parameters) as above, and sampling the parameters given the cluster assignments. The conditional distribution of each cluster's parameters  $\{\theta^{kt}\}$  (for each cluster  $k$ , a sequence of parameters) can be found using Bayes' rule:

$$\mathbb{P}(\{\theta^{kt}\} \mid O, Z) = \mathbb{P}(\{\theta^{kt}\} \mid O|_{z=k}, Z) \propto \left[ \prod_{z_i^t=k} \mathbb{P}(o_i^t \mid \theta^{kt}) \right] \mathbb{P}(\{\theta^{kt}\}) \quad (6)$$

Depending on the choice of parameter priors and observation functions, the resulting conditional distributions can potentially be complicated to represent and difficult to sample from. With additional assumptions that will be presented next, we can find the parameter posterior distribution efficiently and avoid sampling the parameter entirely by “collapsing” it.

### 3.2 Application of DDPs to world modeling

We now apply the DDP mixture model (DDPMM) to our semi-static world modeling problem. For concreteness and simplicity, we consider an instance of the world modeling problem where the fixed attribute  $a$  is the discrete object type (from a finite list of known types), and the dynamic attribute  $x$  is the continuous pose in  $\mathbb{R}^d$  (either 3-D location or 6-D pose). Despite these restrictions, our model and derivations below can be immediately applied to problems with any fixed attributes, and with any dynamic continuous attributes with linear-Gaussian dynamics. Arbitrary dynamic attributes can be represented in our model, but inference will likely be more challenging because in general we will not obtain closed-form expressions.

For our instance of the DDPMM, we assume:

- Time steps in the DDP correspond to epochs in world modeling. This implies that each epoch is modeled as a static DPMM, similar to the problem in Wong et al. (2015) .
- The survival rate only depends on the fixed attribute, i.e.,  $q(\theta) = q(a)$ . (For us, that means the likelihood of object removal is dependent on the object type but not its pose.)
- Likewise, the detection probability only depends on the fixed attribute, i.e.,  $\eta(\theta) = \eta(a)$ .
- The dynamic attribute (pose) follows a random walk with zero-mean Gaussian noise that depends on  $a$  (e.g., a mug likely travels farther per epoch than a table):

$$x^{t+1} = x^t + w, \text{ where } w \sim \mathcal{N}(0, R(a)) \quad (7)$$

This implies that the full transition distribution (of both object type and pose) is:

$$T(\theta^{t+1}; \theta^t) = \mathbb{I}[a^{t+1} = a^t] \mathcal{N}(x^{t+1}; x^t, R(a)) \quad (8)$$

- At each epoch, the DP base distribution has the following form:

$$H(\theta) \triangleq \pi(a) \mathcal{N}(x; \mu^0, \Sigma^0) \quad (9)$$

Here  $\pi$  is a (discrete) prior over the object type, and a normal distribution over the object pose. The initial covariance  $\Sigma^0$  is large, in order to give reasonable likelihood of an object being introduced at any location. In fact, we will set  $\Sigma^0 = \infty I$  and  $\mu^0 = \mathbf{0}$ , representing a noninformative prior over the location. Details can be found in Appendix B.

The above choices for the dynamics and base distribution implies that the parameter posterior and predictive distributions have closed-form expressions. The posterior distribution of the dynamic attribute is a mixture of Gaussians, with a component for each possible value of the fixed attribute  $a$  (since the process noise  $R(a)$  may be different), weighted by the posterior probability of  $a$ . In practice, we track the pose using only the dynamics of the most-likely object type. Thus, in our application, each cluster will maintain a discrete posterior distribution  $\varphi(a)$  for the object type,

and a single Kalman filter / Rauch-Tung-Striebel (RTS) smoother for the object pose distribution. The latter is represented as a sequence of means and covariances  $\{\mu^t, \Sigma^t\}_{t=\xi}^\zeta$  over the cluster's lifetime  $t \in [\xi, \zeta]$ , with the interpretation that  $x^t \sim \mathcal{N}(\mu^t, \Sigma^t)$ .

As mentioned previously, because we have compact representations of the parameter posterior distributions, we can analytically integrate them out sampling them. We first modify the forward sampling equation (Equation 4) to reflect this “collapsing” operation. Since we can no longer condition on the parameters themselves, we instead need to condition on the other observations  $O_{\setminus ti}$  and their current cluster assignments  $Z_{\setminus ti}$ , and use posterior *predictive* likelihoods of the form  $\mathbb{P}(o_i^t | O_{\setminus ti}^k)$  to evaluate the current observation  $o_i^t$ :

$$\begin{aligned} \mathbb{P}(z_i^t = k | o_i^t, O_{\setminus ti}^{\leq t}, Z_{\setminus ti}^{\leq t}) &\propto \mathbb{P}(o_i^t | z_i^t = k, O_{\setminus ti}^{\leq t}, Z_{\setminus ti}^{\leq t}) \mathbb{P}(z_i^t = k | O_{\setminus ti}^{\leq t}, Z_{\setminus ti}^{\leq t}) \\ &\propto \int \left[ \mathbb{P}(o_i^t | \theta^{kt}) \mathbb{P}(\theta^{kt} | O_{\setminus ti}^{k, \leq t}) \right] \mathbb{P}(z_i^t = k | Z_{\setminus ti}^{\leq t}) d\theta^{kt} \\ &\propto \begin{cases} N_{\setminus ti}^{k, \leq t} \int \mathbb{P}(o_i^t | \theta^{kt}) \mathbb{P}(\theta^{kt} | O_{\setminus ti}^{k, \leq t}) d\theta^{kt}, & k \text{ existing, instantiated at } t \\ N_{\setminus ti}^{k, < t} \int \mathbb{P}(o_i^t | \theta^{kt}) \left[ \int \tilde{q}(a^{k\tau}) \tilde{T}(\theta^{kt}; \theta^{k\tau}) \mathbb{P}(\theta^{k\tau} | O_{\setminus ti}^{k, < t}) d\theta^{k\tau} \right] d\theta^{kt}, & k \text{ existing, not instantiated at } t \\ \alpha \int \mathbb{P}(o_i^t | \theta^{kt}) H(\theta^{kt}) d\theta^{kt}, & k \text{ new} \end{cases} \end{aligned} \quad (10)$$

We can now substitute the expressions for  $\mathbb{P}(o_i^t | \theta^{kt})$ ,  $\tilde{T}$ , and  $H$ , where properties of the normal distribution will help us evaluate the integrals. The derivations in Appendix B give the following expressions, as well as details for finding the posterior hyperparameters  $\varphi$ ,  $\mu^{kt}$ , and  $\Sigma^{kt}$  (recall  $\theta^{kt} = (a^k, x^{kt})$ ,  $o_i^t = (b_i^t, y_i^t)$ ):

$$\begin{aligned} \mathbb{P}(z_i^t = k | o_i^t, O_{\setminus ti}^{\leq t}, Z_{\setminus ti}^{\leq t}) &\propto \begin{cases} N_{\setminus ti}^{k, \leq t} \left[ \sum_{a^k} \phi^{a^k}(b_i^t) \varphi(a^k) \right] \mathcal{N}(y_i^t; \mu^{kt}, \Sigma^{kt} + S), & k \text{ existing, instantiated at } t \\ \tilde{q}(\hat{a}^k) N_{\setminus ti}^{k, < t} \left[ \sum_{a^k} \phi^{a^k}(b_i^t) \varphi(a^k) \right] \mathcal{N}(y_i^t; \mu^{k\tau}, \Sigma^{k\tau} + (t - \tau)R(\hat{a}^k) + S), & k \text{ existing, not instantiated at } t \\ \alpha \left[ \sum_{a^k} \phi^{a^k}(b_i^t) \pi(a^k) \right] \text{Unif}(\text{vol}(\text{world})), & k \text{ new} \end{cases} \end{aligned} \quad (11)$$

In the second case, for tractability in filtering, we have assumed that a cluster's dynamics behaves according to its most-likely type  $\hat{a}^k$ ; otherwise, the posterior is a mixture of Gaussians (over all possible transition densities). Also, the third case contains an approximation to avoid evaluating an improper probability density; see Appendix B for details.



## 4 Incorporating World Modeling Constraints

So far, we have only applied a generic DDPMM to our observations, but have ignored the cannot-link constraint, as well as false positives and negatives. We now present modifications to the Gibbs sampler to handle these constraints; the modifications are similar to those from the static case in Wong et al. (2015) .

The cannot-link constraint (see Section 2.2; referred to as “one measurement per object” (OMPO) in Wong et al. (2015) ) couples together cluster assignments for observations within the same view, since we must ensure that no two observations can be assigned to the same existing cluster. For each view, all cluster assignments must be considered together as a joint correspondence vector, and the probability of choosing one such correspondence is proportional to the product of the individual cluster assignment probabilities given in Equation 11. Invalid correspondence vectors that violate the cannot-link constraint are assigned zero probability and hence are not considered; the remaining conditional probabilities are normalized. This can be interpreted as performing *blocked* Gibbs sampling, where blocks are determined by the joint constraints:

$$\mathbb{P}(\mathbf{z}^{tv} \mid \mathbf{o}^{tv}, O_{\setminus tv}, Z_{\setminus tv}) \propto \left[ \prod_i \mathbb{P}(z_i^{tv} \mid o_i^{tv}, O_{\setminus tv}, Z_{\setminus tv}) \right] \mathbb{I}[\mathbf{z}^{tv} \text{ satisfies CLC}] \quad (12)$$

The *correspondence vector*  $\mathbf{z}^{tv}$  is again the concatenation of the individual  $z_i^{tv}$  assignment variables, for all observation indices  $i$  made in view  $v$  at epoch  $t$ ; the interpretation of  $\mathbf{o}^{tv}$  is similar. The individual terms in the product are given by Equation 11 (with the appropriate case depending on the value of  $z_i^{tv}$ ), except now all observations within the same view are excluded (since their assignments are being sampled together) –  $O_{\setminus tv}$  instead of  $O_{\setminus ti}$ , and likewise for assignments  $Z_{\setminus tv}$  and counts  $N_{\setminus tv}$ .

For false positives, we essentially treat it as a special “cluster” that has no underlying parameter. Instead, we assume that if an observation is generated from a false positive, it is generated from some spurious parameter drawn from the base distribution  $H$ , so the likelihood term is the same as that for drawing a new cluster. Like the other cases, we also multiply the likelihood by the number of points already assigned to the cluster, i.e., the number of false positives except for those in the current view. If there are currently no other false positives, then we multiply by the concentration parameter  $\alpha$  instead to ensure that it is always feasible to assign observations to the false positive “cluster”. Also, to incorporate the assumption that false positives are generated with a fixed rate  $\rho$ , we attach a Bernoulli probability to each case in the Gibbs sampler. The false positive conditional probability is multiplied by  $\rho$ , and all other cases are multiplied by  $(1 - \rho)$ . In summary, the conditional probability of an observation being a false positive ( $z = 0$ ) is:

$$\mathbb{P}(z_i^{tv} = 0 \mid o_i^{tv}, O_{\setminus tv}, Z_{\setminus tv}) \propto \left[ \sum_{a^k} \phi^{a^k}(b_i^{tv}) \pi(a^k) \right] \text{Unif}(\text{vol}(\text{world})) \times \begin{cases} \rho N_{\setminus tv}^0, & N_{\setminus tv}^0 > 0 \\ \rho \alpha, & N_{\setminus tv}^0 = 0 \end{cases} \quad (13)$$

The normalizer depends on the other cases in Equation 11 (with additional  $(1 - \rho)$  factors).

Finally, for false negatives, recall that an object that is within the field of view fails to be detected with type-dependent probability  $\eta(a^k)$ . Let  $\delta_k^{tv}$  be 1 if cluster  $k$  is detected in view  $v$  at epoch  $t$ , and 0 otherwise. For a cluster  $k$  that is alive at epoch  $t$  ( $\xi^k \leq t \leq \zeta^k$ ) with parameter  $\theta^{kt}$ ,

the probability of detection is therefore:

$$\mathbb{P}(\delta_k^{tv} = 1) = \left[1 - \eta(a^k)\right] \mathbb{P}(\theta^{kt} \in V^{tv}) = \left[1 - \sum_{a^k} \eta(a^k) \varphi(a^k)\right] \tilde{\Phi}(x^{kt} \in V^{tv}; \mu^{kt}, \Sigma^{kt}) \quad (14)$$

The  $\tilde{\Phi}$  function denotes the CDF of the multivariate normal distribution, with mean  $\mu^{kt}$  and covariance  $\Sigma^{kt}$ . For a particular view  $V^{tv}$ , we only evaluate the above detection probability on clusters that are currently alive at epoch  $t$ . For each such cluster, there is a corresponding  $\delta_k^{tv}$  detection indicator variable, whose value is determined during sampling by the candidate joint correspondence vector  $\mathbf{z}^{tv}$ : if some element of  $\mathbf{z}^{tv}$  is assigned to cluster index  $k$ , then  $\delta_k^{tv} = 1$ ; otherwise,  $\delta_k^{tv} = 0$ . The detection probability for the correspondence vector is:

$$\mathbb{P}_D(\mathbf{z}^{tv} | O_{\setminus tv}, Z_{\setminus tv}) = \prod_{k: \xi^k \leq t \leq \zeta^k} [\mathbb{P}(\delta_k^{tv} = 1)]^{\delta_k^{tv}} [1 - \mathbb{P}(\delta_k^{tv} = 1)]^{1 - \delta_k^{tv}} \quad (15)$$

Putting everything together, we arrive at a constrained blocked collapsed Gibbs sampling inference algorithm. The algorithm takes the observations  $O = \{o_i^{tv}\}$  and visible regions  $\{V^{tv}\}$  as input. As output, the algorithm produces samples from the posterior distribution over correspondence vectors  $\{\mathbf{z}^{tv}\}$ , from which we can compute the posterior parameter distributions  $a^k \sim \varphi$  and  $x^{kt} \sim \mathcal{N}(\mu^{kt}, \Sigma^{kt})$ . The sampling algorithm repeatedly iterates over epochs  $t$  and views  $v$ , each time sampling a new correspondence vector  $\mathbf{z}^{tv}$  from its constrained conditional distribution:

$$\begin{aligned} \mathbb{P}_{\text{View}}(\mathbf{z}^{tv} | \mathbf{o}^{tv}, O_{\setminus tv}, Z_{\setminus tv}) &\propto \left[ \prod_{i: z_i^{tv} \neq 0} (1 - \rho) \mathbb{P}(z_i^{tv} | o_i^{tv}, O_{\setminus tv}, Z_{\setminus tv}) \right] \\ &\times \left[ \prod_{i: z_i^{tv} = 0} \rho \left[ \sum_{a^k} \phi^{a^k}(b_i^{tv}) \pi(a^k) \right] \frac{1}{\text{vol}(\text{world})} \times \begin{cases} N_{\setminus tv}^0, & N_{\setminus tv}^0 > 0 \\ \alpha, & N_{\setminus tv}^0 = 0 \end{cases} \right] \\ &\times \left[ \prod_{k: \xi^k \leq t \leq \zeta^k} [\mathbb{P}(\delta_k^{tv} = 1)]^{\delta_k^{tv}} [1 - \mathbb{P}(\delta_k^{tv} = 1)]^{1 - \delta_k^{tv}} \right] \\ &\times \mathbb{I}[\mathbf{z}^{tv} \text{ satisfies CLC}] \end{aligned} \quad (16)$$

The probability terms in the first and third lines can be found in Equations 11 and 14 respectively.

As in the static case, after incorporating the world modeling constraints, inference becomes inefficient because we now have to compute conditional probabilities for (and sample from) the joint space of correspondence vectors, which in general is exponential in the number of observations in a view. Using the same insights and ideas as in Wong et al. (2015), however, we can adaptively factor the correspondence vector by initially decoupling all assignment variables, then coupling only those that violate the cannot-link constraint.

## 5 Approximate Maximum *a Posteriori* (MAP) inference

We have now presented the entire Gibbs sampling algorithm for DDPMM-based world modeling. However, sampling-based inference can be slow, especially because of the cannot-link constraint

that couples together many latent variables, even if adaptive factoring is used. Although we are interested in maintaining an estimate of our uncertainty in the world, frequently just having the most-likely (maximum *a posteriori* – MAP) world state suffices. In general, even the MAP world model is hard to find (Bar-Shalom and Fortmann, 1988), and many approximate solutions have been proposed.

In the static case, Wong et al. (2015) adapted a hard-clustering algorithm, DP-means, and empirically found that it returned good clustering assignments for some hyperparameter settings. A similar analysis via small-variance asymptotics was performed recently for DDPs, where the mixture components were Gaussian distributions with isotropic noise, resulting in the Dynamic Means algorithm (Campbell et al., 2013). However, there is no simple and principled way to incorporate the additional information from Section 4. Additionally, even without such modifications, the Dynamic Means algorithm requires three free hyperparameters to be specified, which may be significantly harder to tune than the one in DP-means. Instead, we will use a much older idea that does not involve asymptotics, can incorporate all the world-modeling information and constraints, and produces an local optimization algorithm that is similar in spirit to Dynamic Means.

## 5.1 Iterated conditional modes (ICM)

The *iterated conditional modes* (ICM) algorithm performs coordinate ascent on each variable’s conditional distribution, and is guaranteed to converge to a local maximum (Besag, 1986). In particular, instead of iteratively sampling correspondence vectors from their conditional distributions in Gibbs sampling, we find the most-likely one, update parameters based on it, and repeat for each view. Since we are still dealing with the joint space of assignments for all observations in a given view, finding the maximizer still potentially requires searching through a combinatorial space. Fortunately, finding the most-likely correspondence can be formulated as a maximum weighted assignment problem, for which efficient algorithms such as the Hungarian algorithm exist (and have been previously used in data association).

Suppose, for view  $v$  at epoch  $t$ , there are  $M$  observations  $\{o_1, \dots, o_M\}$  and  $K$  existing clusters (possibly not alive/instantiated). Then we wish to match each  $o_i$  to an existing cluster, a new cluster, or a false positive. Any unmatched existing cluster must also be assigned the probability of missed detection. We can solve this as an assignment problem with the following payoff matrix:

	Obs ( $M$ )	FN ( $M + K$ )
Clusters ( $K$ )	$\log \mathbb{P}(z_i = k) + \log(1 - \rho)$ $+ \mathbb{I}[\xi^k \leq t \leq \zeta^k] \log \mathbb{P}(\delta_k = 1)$	$\mathbb{I}[\xi^k \leq t \leq \zeta^k] \log \mathbb{P}(\delta_k = 0)$
New ( $M$ )	$\log \mathbb{P}(z_i = \text{new}) + \log(1 - \rho)$	0
FP ( $M$ )	$\log \mathbb{P}(z_i = 0 \text{ (FP)}) + \log \rho$	0

The payoff matrix has  $2M + K$  entries (indicated in parentheses), to allow for the case that all observations are assigned to new clusters, and likewise that all are spurious. Any extra New/FP nodes are assigned to extra FN nodes, with zero payoff. The payoffs in the first column are: for an existing cluster, given by cases 1 and 2 in Equation 11, depending on whether or not the cluster has been instantiated yet; for a new cluster, given by case 3 in Equation 11; and for a false positive, given by Equation 13. Note that log probabilities are used to decompose the view’s joint correspondence probability into a sum of individual terms. By construction, the cannot-link constraint is satisfied.

**Input:** Observations  $O = \{o_i^{tv}\}$

Visible regions  $\{V^{tv}\}$

Number of samples  $N$

**Output:** Samples of cluster assignments  $\{\mathbf{z}^{tv}\}$

- 1: Init. all entries to  $-1$  (FP) in  $Z^{(0)} = \{\mathbf{z}^{tv}\}^{(0)}$
- 2: **for**  $n := 1$  **to**  $N$  **do**
- 3:    $Z' := \text{Proposal}(Z^{(n-1)})$  (see Oh et al. (2009))
- 4:    $A(Z^{(n-1)} \rightarrow Z') :=$   
 $\min \left( 1, \frac{\text{Likelihood}(Z') \mathbb{P}(Z' \rightarrow Z^{(n-1)})}{\text{Likelihood}(Z^{(n-1)}) \mathbb{P}(Z^{(n-1)} \rightarrow Z')} \right)$
- 5:   Sample  $u \sim \text{Unif}(0, 1)$
- 6:   **if**  $u < A(Z^{(n-1)} \rightarrow Z')$  **then**
- 7:      $Z^n = Z'$
- 8:   **else**
- 9:      $Z^n = Z^{(n-1)}$

(a) MCMCDA (Oh et al., 2009) for DDPMM

**Input:** Observations  $O = \{o_i^{tv}\}$

Visible regions  $\{V^{tv}\}$

Number of samples  $N$

**Output:** Samples of cluster assignments  $\{\mathbf{z}^{tv}\}$

- 1: Init. all entries to  $-1$  (FP) in  $Z^{(0)} = \{\mathbf{z}^{tv}\}^{(0)}$
- 2: **repeat**
- 3:   **for**  $t := 1$  **to**  $T$ ;  $v := 1$  **to**  $V^t$  **do**
- 4:     Solve ICM weighted assignment problem  
for most-likely  $\mathbf{z}^{tv}$ , given  $Z_{\setminus v}^t$
- 5:   **until** convergence
- 6:   Construct new dataset  $C = \{c_i^t\}$  with  
a single data point for each non-FP cluster  
found by ICM (at the same epoch)
- 7:   Sample tracks  $L$  by performing MCMCDA on  $C$
- 8:   Convert track samples to cluster assignments

(b) Two-stage inference algorithm for DDPMM

**Figure 2:** Two algorithms for performing inference in DDPMMs, one by Metropolis-Hastings (MH) (Oh et al., 2009), the other a two-stage procedure involving ICM, followed by the MH procedure.

## 5.2 A two-stage inference scheme

Although the ICM algorithm presented can find good clusters at a single epoch very quickly, we will see in experiments that it does not converge to good cluster trajectories. The issue is that ICM moves are local, in that it considers one view at a time. Suppose we have identified correctly all objects in epoch 1 using ICM. When we consider the first view in epoch 2, there may be significant changes present, and using the first view only, ICM decides whether or not to assign the new observations to existing clusters (by reviving them from the previous epoch). Since the uncertainty in the object states immediately after a transition is high, basing the cluster connectivity decisions on a single view is unreliable.

This suggests a two-level inference scheme. Since ICM can reliably find good clusters within single epochs, we first apply ICM to each epoch’s data *independently*, treating them as unrelated static worlds. Next, we attempt to connect clusters between different epochs. This is essentially another tracking problem, although the likelihood function is somewhat different (depends on many underlying data points), and is much reduced in size. Since the problem is significantly smaller, traditional tracking methods such as MHT can be applied to this cluster-level tracking problem.

We present one such scheme in Algorithm 2(b), using MCMCDA (Algorithm 2(a); Oh et al. (2009)) to solve the cluster-level problem. We choose a batch-mode sampling algorithm such as MCMCDA because it can return samples from the posterior distribution, and has an attractive anytime property – we can terminate at any point and still return a list of valid samples. For inferring the MAP configuration, the best sample can be returned instead. Since we are sampling from the true posterior distribution, in the limit of infinite samples, the true MAP configuration will be found almost surely.

To apply MCMCDA, we need to evaluate the likelihood of a complete configuration  $Z$ , encompassing all epochs and views (line 4 in Algorithm 2(a)). To do so, we first find the posterior parameter distributions for the clusters/objects (as given by  $Z$ ) using Appendix B, then combine

the observation likelihoods (Equation 36), as well as the false positive and false negative priors:

$$\begin{aligned}
\mathbb{P}(O|Z)\mathbb{P}(Z) &= \prod_t \prod_v \mathbb{P}(\mathbf{o}^{tv} | \mathbf{z}^{tv}) \mathbb{P}_{\text{FP}}(\mathbf{z}^{tv}) \mathbb{P}_{\text{FN}}(\mathbf{z}^{tv}) \\
&= \prod_t \prod_v \left\{ \left[ \prod_i \int \mathbb{P}(o_i^t | \theta^{kt}, z_i^t = k) \mathbb{P}(\theta^{kt}) d\theta^{kt} \right] \right. \\
&\quad \times \text{Bin}(N_{z=0}^{tv} | N^{tv}, \rho) \\
&\quad \times \left. \left[ \prod_{k: \xi^k \leq t \leq \zeta^k} [\mathbb{P}(\delta_k^{tv} = 1)]^{\delta_k^{tv}} [1 - \mathbb{P}(\delta_k^{tv} = 1)]^{1-\delta_k^{tv}} \right] \right\} \quad (17)
\end{aligned}$$

## 6 Experiments

Approximate MAP inference for world modeling via ICM, MCMCDA, and the two-stage algorithm ICM-MCMC were tested on a simulated domain, and also on a sequence of real robot vision data constructed from the static scenes in Wong et al. (2015). To perform MAP inference on MCMCDA and ICM-MCMC, the most-likely sample (as scored by Equation 17) was chosen, from  $10^5$  samples in MCMCDA, and  $10^4$  in the second stage of ICM-MCMC. In both experiments, ICM-MCMC significantly outperforms the other two methods, and even ICM performs better than MCMCDA.

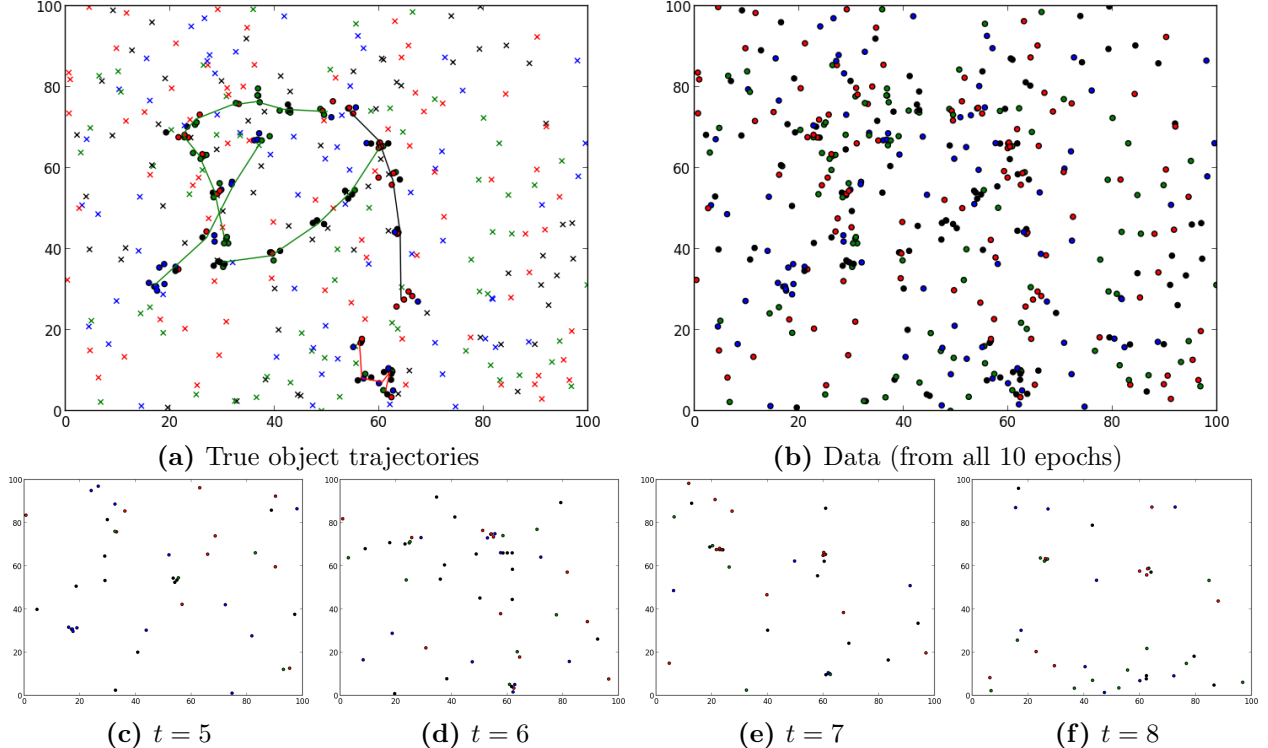
### 6.1 Simulation

Objects in our simulated domain had one of four fixed object types, a time-evolving location  $(x, y) \in [0, 100] \times [0, 100]$ , and a time-evolving velocity vector. Observations were made in 10 epochs of this domain, with 5 views per epoch (visible region is the entire domain). In total, 5 objects existed, each for some contiguous sub-interval of the elapsed time. Within each view, the number of false positives was generated from  $\text{Poi}(5)$ , and the probability of a missed detection was 0.1. The correct object type was observed with probability 0.6, with equal likelihood (0.1) of being confused with the other 3 object types. Locations were observed with isotropic Gaussian noise, standard deviation 1.0. The object’s velocity vector was maintained from the previous time step, with added Gaussian noise, standard deviation 5.0. Between epochs, the probability of survival was 0.9. The observed data (i.e., the algorithm input) and the true object states are shown in Figure 3.

The resulting MAP clusters found by ICM, MCMCDA, and ICM-MCMC are shown in Figure 4, along with their log-likelihood values (higher / less negative is better). ICM-MCMC clearly outperforms the other methods, and finds essentially the same clusters as given by the true association. The clusters found generally have tight covariance values, unlike those in ICM and MCMCDA. These two methods, especially MCMCDA, tend to find many more clusters than are truly present.

### 6.2 Using robot data from static scenes

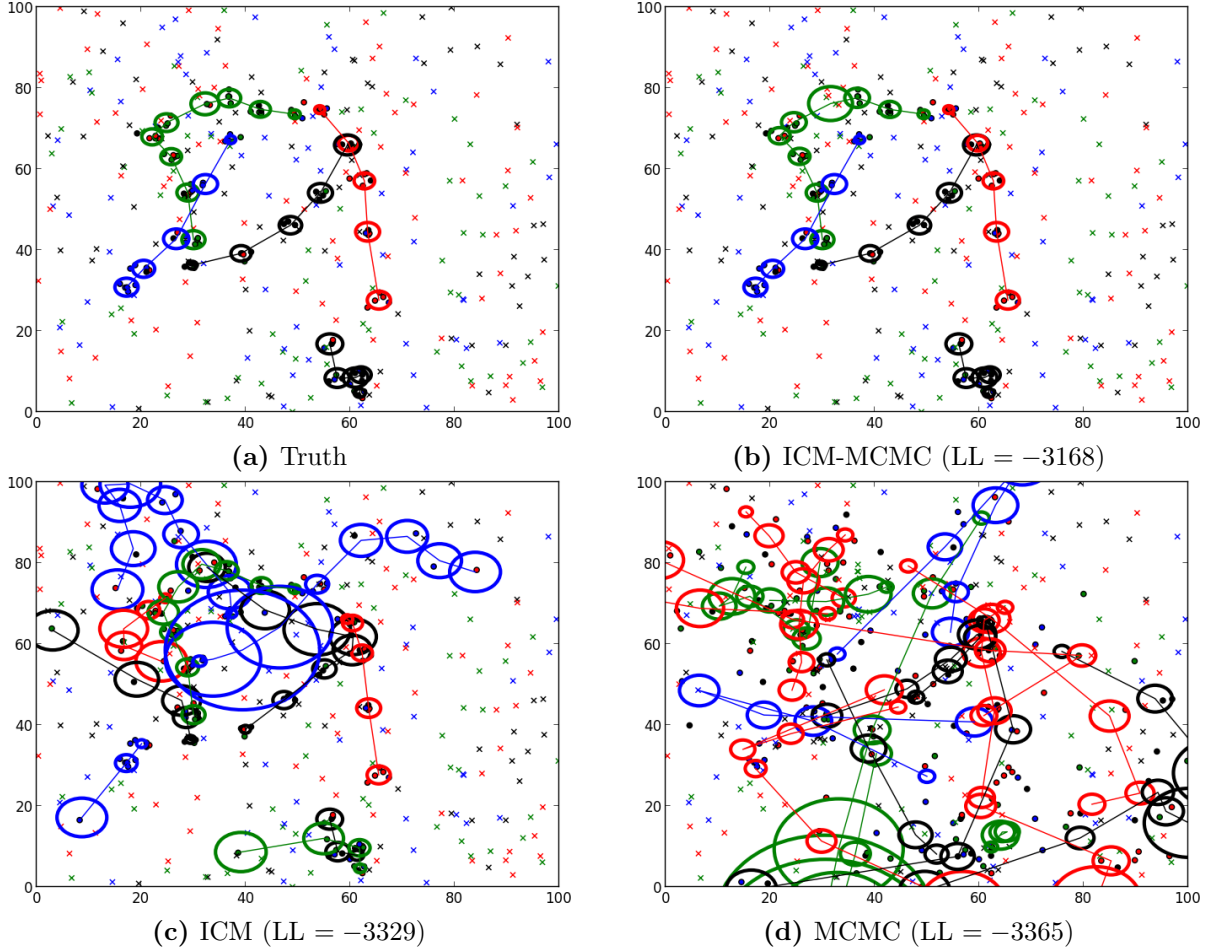
We also applied the same algorithms to the static robot vision data that were used in Wong et al. (2015) to evaluate DPMM methods. To convert static scenes into dynamic scenes, we choose static scenes that were reasonably similar, and simply concatenated their data together, as if each scene corresponded to a different epoch. One such example is shown in Figure 5.



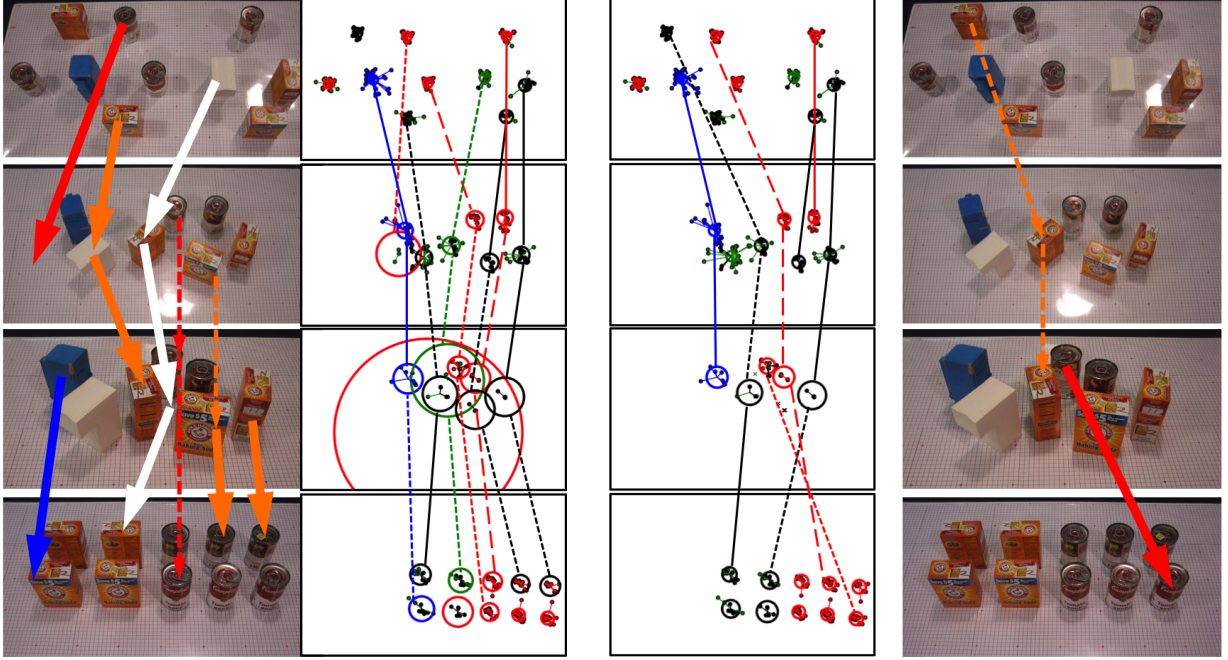
**Figure 3:** Data and object states in a simulated domain. The top left shows the true object  $(x, y)$  locations and their trajectories over time, color-coded by their associated object type. Observations are shown as filled dots (corresponding to true positives) and crosses (false positives). The top right shows the data from all 10 epochs (5 views per epoch) that is given as input, without any information about the underlying object states and associations. Some form of clustering over views and time is visible. A more realistic view of the data is shown in the bottom row, for a sequence of 4 epochs.

Objects in different scenes were all placed on the same tabletop of dimensions  $1.2\text{m} \times 0.6\text{m}$ ; all data were placed in the table’s frame of reference. Four object types were present, and typically each scene had 5–10 objects. Unlike the previous simulation, we do not assume objects have velocities; between epochs, we assume that the location changes with isotropic Gaussian noise, standard deviation 0.1. Since changes were significant between epochs, we assumed a relatively low 0.5 probability of survival. Object locations are sensed with Gaussian noise, standard deviation 0.03; the object type noise model and probability of detection is the same as before. The probability of false positives is much lower for this domain; we assumed the number of false positives had a  $\text{Poi}(0.1)$  distribution.

Figure 5 shows the MAP associations found by ICM and ICM-MCMC, with lines connecting cluster states over epochs. Annotations were also added (in the form of three different line styles) to facilitate comparison between the ICM and ICM-MCMC results; see figure caption for details. ICM tends to suggest many more transitions than ICM-MCMC, many of which are actually implausible.



**Figure 4:** The clusters found for the simulated domain are shown in thick ellipses, centered at the location mean, color-coded by the most-likely object type inferred (across the entire trajectory, since it is a static attribute). The ellipses depict a level set of the posterior location distribution (uncertainty given by Gaussian covariance matrix). The posterior clusters derived from the true association is shown in the top left; the one found by ICM-MCMC is essentially identical (with a minor difference in the green track). In contrast, the posterior clusters found by ICM and the most-likely sample from MCMC (of  $10^5$ ), shown in the bottom row, are qualitatively much different, and have significantly lower log-likelihood (LL) values.



(a) ICM transitions not present in ICM-MCMC (b) Most-likely ICM configuration (LL = -968) (c) Most-likely ICM-MCMC conf (LL = -931) (d) ICM-MCMC transitions not present in ICM

**Figure 5:** Approximate MAP cluster (object) trajectories found using ICM and ICM-MCMC on the robot vision data collection in Wong et al. (2015). The concatenated sequence of scenes (epochs) is shown from top to bottom. The inferred clusters and tracks are shown in the middle two columns. Lines connecting cluster pairs between epochs are color-coded by the inferred object type (fixed across epochs), and are marked by one of three line styles used to compare results from the two algorithms. A solid line means the same pair was connected by both algorithms; a dashed line means a similar pair (in likelihood) was connected; a dotted line means the pair was not connected by the other algorithm. To make the differences clearer, the top-down reference views have been annotated with arrows, for pairs of objects that were only connected by one algorithm (dotted lines in the middle two). The left column shows pairs that were connected by ICM but not ICM-MCMC; the right column shows the opposite. Solid arrows depict transitions that are unlikely, whereas dashed arrows depict plausible transitions. ICM tends to suggest many more transitions than ICM-MCMC, many of which are actually implausible.



## A Background on dependent Dirichlet processes

Lin et al. (2010) exploited the fact that there exists a one-to-one correspondence between DPs over space  $\Omega$  and spatial Poisson processes in the product space  $\Omega \times \mathbb{R}_+$ . This means that an underlying Poisson process can be extracted from any DP, and vice versa. By considering transitions on the underlying Poisson processes, and restricting to transition steps where the Poisson process remains closed under transition (more fundamentally, by preserving complete randomness), we obtain a new spatial Poisson process at the next time step, which can be converted back to a new DP.

According to the stick-breaking construction of the DP (Sethuraman, 1994), if  $D^t \sim \text{DP}$ , then it can be expressed as infinite sum of weighted atoms:  $D^t = \sum_{i=1}^{\infty} w_i \delta_{\theta_i}$ , where  $w_i \in \mathbb{R}_+$ , and  $\theta_i \in \Omega$ . Then the following DP-preserving transition steps are applied in order:

- **Subsampling (removal):** Let  $q : \Omega \rightarrow [0, 1]$  be a parameter-dependent survival rate, i.e.,  $q(\theta)$  specifies how likely some  $\theta$  in the current time step survives in the next time step. For each atom  $\theta_i$ , draw  $b_i \sim \text{Ber}(q(\theta_i))$ , and retain atoms with  $b_i = 1$ . Renormalizing the weights on the retained atoms gives a new DP  $D' = \sum_{i:b_i=1} w'_i \delta_{\theta_i}$  (where  $\sum_{i:b_i=1} w'_i = 1$ ).
- **Point transition (movement):** Let  $T(\cdot; \theta) : \Omega \rightarrow \mathbb{R}_+$  be a parameter-dependent transition function, i.e.,  $T(\theta'; \theta)$  specifies how likely some  $\theta$  in the current time step moves to  $\theta'$  in the next time step, given that it survives. For each atom  $\theta_i$ , draw  $\theta'_i \sim T(\cdot; \theta)$ . Then  $D'' = \sum_{i:b_i=1} w'_i \delta_{\theta'_i}$  is a new DP.
- **Superposition (addition):** Let  $\Delta = \sum_j \varpi_j \delta_{\vartheta_j}$  be a new independent DP, and let  $(c, d) \sim \text{Dir}(\alpha'', \alpha)$ , where  $\alpha''$  and  $\alpha$  are the concentration parameters of  $D''$  and  $\Delta$  respectively. Then the random convex combination  $D^{t+1} = cD'' + d\Delta$  is a DP, and acts as the prior for the next time step.

The upshot of this DDP construction is that, if we marginalize out the DP prior, we get the following prior for  $\theta^{t+1}$ , given the parameters from the previous time  $\Theta^t$ :

$$\theta^{t+1} \mid \Theta^t \propto \alpha H(\theta^{t+1}) + \sum_k q(\theta^{kt}) N^{k, \leq t} T(\theta^{t+1}; \theta^{kt}) \quad (18)$$

The first term is for new atoms, drawn from a DP with base distribution  $H(\theta)$  and concentration parameter  $\alpha$ .<sup>2</sup> The second term corresponds to existing atoms that have undergone subsampling and transition steps; these steps affect the assignment probability, as indicated by the presence of  $q$  and  $T$ . Additionally,  $N^{k, \leq t}$  is the number of points that have been assigned to cluster  $k$ , for all time steps up to time  $t$ . This term is similar to that in the DP. Notice that if  $q \equiv 1$  and  $T(\cdot; \theta^{kt}) = \delta_{\theta^{kt}}$ , then we exactly get back the predictive distribution in the DP.

Since  $\theta^{t+1} \sim D^{t+1}$ , and  $D^{t+1}$  is a DP, we can find the predictive distribution of  $\theta^{t+1}$ , conditioning also on parameters  $\Theta^{t+1}$  that have been instantiated at time  $(t+1)$ :

$$\theta^{t+1} \mid \Theta^t \propto \alpha H(\theta^{t+1}) + \sum_{k: N^{k, t+1} > 0} N^{k, \leq t+1} \mathbb{I}[\theta^{k, t+1} = \theta^{kt}] + \sum_{k: N^{k, t+1} = 0} q(\theta^{kt}) N^{k, \leq t} T(\theta^{t+1}; \theta^{kt}) \quad (19)$$

---

<sup>2</sup>Technically,  $\alpha$  includes both the innovation process from the superposition step, as well as a subsampled and transitioned version of innovation processes from previous times; see Lin (2012).

In general, some atoms may not be observed for several time steps, but still affect the prior (with decayed weight and dispersed parameter values). Also, some clusters may already have been instantiated at the current time  $t$ , either newly drawn from the innovation process, or transitioned from existing atoms. The general form of the prior on  $\theta^t$  is:

$$\theta^t \mid \Theta^{\leq t} \propto \alpha H(\theta^t) + \sum_{k: N^{kt}=0} q^{kt} N^{k, \leq \tau} T(\theta^t; \theta^{k\tau}) + \sum_{k: N^{kt}>0} N^{k, \leq t} \mathbb{I}[\theta = \theta^{kt}] \quad (20)$$

The first two terms are similar to those in Equation 18 above, except the sum is only over clusters that have not been instantiated at time  $t$  ( $N^{kt} = 0$ ). These existing clusters may be ‘revived’, but they are weighted by accumulated subsampling and transition terms, based on the previous time  $\tau = \tau^{kt}$  at which they were instantiated:

$$q^{kt} \triangleq \left[ q(\theta^{k\tau}) \right]^{t-\tau} \quad (21)$$

$$T(\theta^t; \theta^{k\tau}) \triangleq \int \dots \int \prod_{t'=\tau+1}^t T(\theta^{t'}; \theta^{t'-1}) d\theta^{\tau+1} \dots d\theta^{t-1}$$

The third term in Equation 20 corresponds to atoms that have been instantiated at the current time  $t$ . In this case, we know both that the atom survived and its current value, so  $q$  and  $T$  disappear. Also, the count  $N^{k, \leq t}$  now includes cluster assignments at the current time  $t$  as well.

## B Derivation of closed-form inference expressions for application of DDPs to world modeling (Section 3.2)

In this appendix, we derive closed-form expressions for the posterior and predictive distributions of the parameter  $\theta = (a, x)$ , under the assumptions specified in Section 3.2.

The expressions for the fixed attribute are the same as in Wong et al. (2015), since it is static. For convenience, we reproduce the equations here. Given a set of observations  $\{b\}$ :

$$\varphi(a) \triangleq \mathbb{P}(a \mid \{b\}) \propto \mathbb{P}(\{b\} \mid a) \mathbb{P}(a) \propto \left[ \prod_{b_i \in \{b\}} \phi^a(b_i) \right] \pi(a) \quad (22)$$

$$\mathbb{P}(b' \mid \{b\}) \propto \sum_a \mathbb{P}(b' \mid a) \mathbb{P}(a \mid \{b\}) = \sum_a \phi^a(b') \varphi(a) \quad (23)$$

Given a set of observations  $\left\{ \left\{ y_i^t \right\}_{i=1}^{N^t} \right\}_{t=\xi}^{\zeta}$  of the dynamic attributes, we can find the posterior distribution on  $\{x^t\}_{t=\xi}^{\zeta}$  by performing Kalman filtering and smoothing. Applying a generic Kalman filter to the world modeling problem gives the following recursive filtering equations for  $(\tilde{\mu}, \tilde{\Sigma})$ , the hyperparameters in the forward direction (during filtering):

$$\begin{aligned} \hat{\mu}^t &= \tilde{\mu}^{t-1}, & \hat{\Sigma}^t &= \tilde{\Sigma}^{t-1} + R(a) \\ K^t &= \begin{cases} \hat{\Sigma}^t \left( \hat{\Sigma}^t + \frac{S}{N^t} \right)^{-1}, & N^t > 0 \\ \mathbf{0}, & N^t = 0 \end{cases} \\ \tilde{\mu}^t &= \hat{\mu}^t + K^t (\bar{y}^t - \hat{\mu}^t), & \tilde{\Sigma}^t &= (I - K^t) \hat{\Sigma}^t \end{aligned} \quad (24)$$

Recall that  $R(a)$  is the covariance per time step of the random walk on  $x$ , and  $S$  is the covariance of the measurement noise distribution. The “ $\hat{\cdot}$ ” variables are the predicted parameters before incorporating observations, and the “ $\tilde{\cdot}$ ” variables are the parameters after incorporating observations (i.e., the Kalman filter output). Since there may be multiple observations of the pose in a single epoch, we have used an equivalent formulation involving the sample means  $\bar{y}$ , by exploiting the fact that if each  $y_i^t \sim \mathcal{N}(x^t, S)$ , then the sample mean has distribution  $\bar{y}^t \sim \mathcal{N}(x^t, \frac{S}{N^t})$ . There may also be no observations at a given time, in which case the correction step has no effect ( $K^t = 0$ ).

The Kalman filter is initialized with a noninformative prior:

$$\mu^0 = \mathbf{0}, \Sigma^0 = \infty I \quad (25)$$

In practice, this implies that after the initial measurement(s) at time  $\xi$ ,  $\tilde{x}^\xi \sim \mathcal{N}(\bar{y}^\xi, \frac{S}{N^\xi})$ . To see this, we can apply Equation 24 on  $(\mu^0, \Sigma^0)$ :

$$K^\xi = (\Sigma^0 + R(a)) \left( \Sigma^0 + R(a) + \frac{S}{N^\xi} \right)^{-1} \quad (26)$$

$$\tilde{\mu}^\xi = \mu^0 + K^\xi (\bar{y}^\xi - \mu^0) = K^\xi \bar{y}^\xi \quad (27)$$

$$\tilde{\Sigma}^\xi = (I - K^\xi) (\Sigma^0 + R(a)) \quad (28)$$

To handle the infinite initial covariance, we interpret  $\Sigma^0$  as  $\lim_{n \rightarrow \infty} nI$ . This leads to:

$$\begin{aligned} K^\xi &= \lim_{n \rightarrow \infty} \left[ nI \left( nI + R(a) + \frac{S}{N^\xi} \right)^{-1} + R(a) \left( nI + R(a) + \frac{S}{N^\xi} \right)^{-1} \right] \\ &= \lim_{n \rightarrow \infty} \left[ \left( I + \frac{R(a)}{n} + \frac{1}{n} \frac{S}{N^\xi} \right)^{-1} + \frac{1}{n} R(a) \left( I + \frac{R(a)}{n} + \frac{1}{n} \frac{S}{N^\xi} \right)^{-1} \right] \\ &= I + 0 \cdot R(a) \cdot I = I \end{aligned} \quad (29)$$

Hence  $\tilde{\mu}^\xi = K^\xi \bar{y}^\xi = \bar{y}^\xi$ . For the covariance:

$$\begin{aligned} \tilde{\Sigma}^\xi &= \left[ I - (\Sigma^0 + R(a)) \left( \Sigma^0 + R(a) + \frac{S}{N^\xi} \right)^{-1} \right] (\Sigma^0 + R(a)) \\ &= \left[ \left( \Sigma^0 + R(a) + \frac{S}{N^\xi} \right) \left( \Sigma^0 + R(a) + \frac{S}{N^\xi} \right)^{-1} - (\Sigma^0 + R(a)) \left( \Sigma^0 + R(a) + \frac{S}{N^\xi} \right)^{-1} \right] (\Sigma^0 + R(a)) \\ &= \frac{S}{N^\xi} \left( \Sigma^0 + R(a) + \frac{S}{N^\xi} \right)^{-1} (\Sigma^0 + R(a)) \\ &= \lim_{n \rightarrow \infty} \left[ \frac{S}{N^\xi} \left( nI + R(a) + \frac{S}{N^\xi} \right)^{-1} nI + \frac{S}{N^\xi} \left( nI + R(a) + \frac{S}{N^\xi} \right)^{-1} R(a) \right] \\ &= \lim_{n \rightarrow \infty} \left[ \frac{S}{N^\xi} \left( I + \frac{R(a)}{n} + \frac{1}{n} \frac{S}{N^\xi} \right)^{-1} + \frac{1}{n} \frac{S}{N^\xi} \left( I + \frac{R(a)}{n} + \frac{1}{n} \frac{S}{N^\xi} \right)^{-1} R(a) \right] \\ &= \frac{S}{N^\xi} \cdot I + 0 \cdot \frac{S}{N^\xi} \cdot I \cdot R(a) = \frac{S}{N^\xi} \end{aligned} \quad (30)$$

In summary, choosing  $(\mu^0, \Sigma^0) = (\mathbf{0}, \infty I)$  is equivalent to initializing the Kalman filter with  $(\mu^\xi, \Sigma^\xi) = (\bar{y}^\xi, \frac{S}{N^\xi})$ , and proceeding for times  $\xi < t \leq \zeta$ .

After proceeding forward in time, information from later observations should also be propagated *backward* in time via a smoothing operation. For example, for our application, the Rauch-Tung-Striebel (RTS) smoother runs the following recursive operations starting at time  $\zeta$ :

$$\mu^\zeta = \tilde{\mu}^\zeta, \quad \Sigma^\zeta = \tilde{\Sigma}^\zeta \quad (31)$$

$$C^t = \tilde{\Sigma}^t \left( \hat{\Sigma}^{t+1} \right)^{-1} = \tilde{\Sigma}^t \left( \tilde{\Sigma}^t + R(a) \right)^{-1} \quad (32)$$

$$\mu^t = \tilde{\mu}^t + C^t \left( \mu^{t+1} - \hat{\mu}^{t+1} \right) = \tilde{\mu}^t + C^t \left( \mu^{t+1} - \tilde{\mu}^t \right) \quad (33)$$

$$\Sigma^t = \tilde{\Sigma}^t + C^t \left( \Sigma^{t+1} - \hat{\Sigma}^{t+1} \right) (C^t)^\top = \tilde{\Sigma}^t + C^t \left( \Sigma^{t+1} - \tilde{\Sigma}^t - R(a) \right) (C^t)^\top \quad (34)$$

Recall that “ $\hat{\cdot}$ ” and “ $\tilde{\cdot}$ ” variables are the predicted and filtered parameters respectively. Parameters without such modifications are smoothed.

Once the sequence of parameters  $\{\mu^t, \Sigma^t\}_{t=\xi}^\zeta$  is inferred, we can use them to determine the log-likelihood of the observations (for scoring associations) and the predictive distributions (for determining cluster assignment in Gibbs sampling). We will repeatedly use the following fact:

$$x \sim \mathcal{N}(\mu, \Sigma), \quad y|x \sim \mathcal{N}(x, \Lambda) \Rightarrow y \sim \int \mathbb{P}(y|x) \mathbb{P}(x) dx = \mathcal{N}(\mu, \Sigma + \Lambda) \quad (35)$$

For example, we know that  $x^t \sim \mathcal{N}(\mu^t, \Sigma^t)$  (hyperparameters obtained from Kalman smoothing), and from our modeling assumptions,  $y^t|x^t \sim \mathcal{N}(x^t, S)$ . Hence the marginal distribution over the pose observation (marginalized over all possible latent poses  $x^t$ ) is  $y^t \sim \mathcal{N}(\mu^t, \Sigma^t + S)$ . From this we can immediately find the marginal likelihood of the observed data:

$$\mathbb{P} \left( \left\{ \{y_i^t\}_{i=1}^{N^t} \right\}_{t=\xi}^\zeta \right) = \prod_{t=\xi}^\zeta \prod_{i=1}^{N^t} \mathbb{P}(y_i^t) = \prod_{t=\xi}^\zeta \prod_{i=1}^{N^t} \mathcal{N}(y_i^t; \mu^t, \Sigma^t + S) \quad (36)$$

This likelihood expression is used to score potential association hypotheses.

We can now derive the conditional probability expressions in the Gibbs sampler, shown in Equation 10. In collapsed Gibbs sampling, each observation’s predictive likelihood  $\mathbb{P}(o_i^t | O_{\setminus ti}^k)$  involves an integral over the latent parameters  $\theta^{kt} = (a^k, x^{kt})$  of the cluster. In forward sampling, assigning observation  $o_i^t$  to cluster  $k$  has three cases:

1. If cluster  $k$  exists and is instantiated (i.e., has other observations at time  $t$  assigned to it), the posterior distribution of the pose  $x^{kt}$  is  $\mathcal{N}(\mu^{kt}, \Sigma^{kt})$ , and the posterior distribution of the fixed attribute is  $\varphi(a^k)$ . Thus the predictive distribution is:

$$\begin{aligned} \mathbb{P}(o_i^t | O_{\setminus ti}^k) &= \int \mathbb{P}(o_i^t | \theta) \mathbb{P}(\theta | O_{\setminus ti}^k) d\theta \\ &= \left[ \sum_{a^k} \mathbb{P}(b_i^t | a^k) \varphi(a^k) \right] \int \mathbb{P}(y_i^t | x^{kt}) \mathcal{N}(x^{kt}; \mu^{kt}, \Sigma^{kt}) dx^{kt} \\ &= \left[ \sum_{a^k} \phi^{a^k}(b_i^t) \varphi(a^k) \right] \mathcal{N}(y_i^t; \mu^{kt}, \Sigma^{kt} + S) \end{aligned} \quad (37)$$

In the final line, we used Equations 23 and 35 to simplify the predictive distributions.

2. If cluster  $k$  exists, but it has not yet been instantiated, this implies that, in the forward case, that the time of the observation  $t$  is beyond the final observed time  $\tau = \zeta$  associated with the cluster. Then instead of integrating over the posterior distribution of  $x^{kt}$ , which does not exist yet, we need to integrate over its *predictive* distribution. This can be found by propagating the prediction step in the Kalman filter, starting from the final time step's distribution  $x^{k\tau} \sim \mathcal{N}(\mu^{k\tau}, \Sigma^{k\tau})$ :

$$\mu^{kt} = \mu^{k\tau}, \quad \Sigma^{kt} = \Sigma^{k\tau} + (t - \tau)R(a^k) \quad (38)$$

This is precisely the “generalized” transition distribution  $\tilde{T}$  for the pose in the DDPMM. Hence  $x^{kt} \sim \mathcal{N}(\mu^{k\tau}, \Sigma^{k\tau} + (t - \tau)R(a^k))$ , and by a derivation similar to Equation 37:

$$\mathbb{P}(o_i^t | O_{\setminus ti}^k) = \left[ \sum_{a^k} \phi^{a^k}(b_i^t) \varphi(a^k) \right] \mathcal{N}(y_i^t; \mu^{k\tau}, \Sigma^{k\tau} + (t - \tau)R(a^k) + S) \quad (39)$$

3. If cluster  $k$  does not exist (and  $O_{\setminus ti}^k = \emptyset$ ), then we should use the base distribution  $H(\theta) \triangleq \pi(a) \mathcal{N}(x; \mu^0, \Sigma^0)$  instead of the posterior distribution. Then:

$$\begin{aligned} \mathbb{P}(o_i^t | O_{\setminus ti}^k) &= \mathbb{P}(o_i^t) = \int \mathbb{P}(o_i^t | \theta) H(\theta) d\theta \\ &= \left[ \sum_{a^k} \mathbb{P}(b_i^t | a^k) \pi(a^k) \right] \int \mathbb{P}(y_i^t | x^{kt}) \mathcal{N}(x^{kt}; \mu^0, \Sigma^0) dx^{kt} \\ &= \left[ \sum_{a^k} \phi^{a^k}(b_i^t) \pi(a^k) \right] \mathcal{N}(y_i^t; \mathbf{0}, \infty I) \end{aligned} \quad (40)$$

However, this requires the evaluation of an improper normal distribution in the final term. Since the choice of this prior was motivated by an attempt to give all initial poses equal probability, the same effect can be achieved by using a uniform distribution over the total explored world volume. Thus, in practice during Gibbs sampling we evaluate the following:

$$\mathbb{P}(o_i^t | O_{\setminus ti}^k) = \left[ \sum_{a^k} \phi^{a^k}(b_i^t) \pi(a^k) \right] \text{Unif}(\text{vol}(\text{world})) \quad (41)$$

Note that this is similar to the expression for the observation likelihood of false positives. If the observation is actually assigned to a new cluster, then we revert to the noninformative normal prior and perform Kalman smoothing, which is now no longer problematic since it does not require evaluation of improper densities.

## References

- A. Ahmed and E. Xing. Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: with applications to evolutionary clustering. In *SIAM International Conference on Data Mining*, 2008.
- C.E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.
- Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association*. Academic Press, 1988.
- J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B*, 48:259–302, 1986.
- T. Campbell, M. Liu, B. Kulis, J.P. How, and L. Carin. Dynamic clustering via asymptotics of the dependent Dirichlet process mixture. In *Advances in Neural Information Processing Systems*, 2013.
- I.J. Cox and J.J. Leonard. Modeling a dynamic environment using a Bayesian multiple hypothesis approach. *Artificial Intelligence*, 66(2):311–344, 1994.
- F. Dellaert, S.M. Seitz, C.E. Thorpe, and S. Thrun. EM, MCMC, and chain flipping for structure from motion with unknown correspondence. *Machine Learning*, 50(1–2):45–71, 2003.
- J. Elfring, S. van den Dries, M.J.G. van de Molengraft, and M. Steinbuch. Semantic world modeling using probabilistic multiple hypothesis anchoring. *Robotics and Autonomous Systems*, 61(2):95–105, 2013.
- D. Lin. *Generative Modeling of Dynamic Visual Scenes*. PhD thesis, Massachusetts Institute of Technology, 2012.
- D. Lin, E. Grimson, and J. Fisher. Construction of dependent Dirichlet processes based on Poisson processes. In *Advances in Neural Information Processing Systems*, 2010.
- S.N. MacEachern. Dependent nonparametric processes. In *ASA Section on Bayesian Statistics*, 1999.
- S.N. MacEachern. Dependent Dirichlet processes. Technical report, Ohio State University, 2000.
- R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.
- S. Oh, S. Russell, and S. Sastry. Markov chain Monte Carlo data association for multi-target tracking. *IEEE Transactions on Automatic Control*, 54(3):481–497, 2009.
- H. Pasula, S. Russell, M. Ostland, and Y. Ritov. Tracking many objects with many sensors. In *International Joint Conference on Artificial Intelligence*, 1999.
- D.B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6):843–854, 1979.
- J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

- Y.W. Teh. Dirichlet processes. In C. Sammut and G.I. Webb, editors, *Encyclopedia of Machine Learning*, pages 280–287. Springer, 2010.
- L.L.S. Wong, L.P. Kaelbling, and T. Lozano-Pérez. Data association for semantic world modeling from partial views. *International Journal of Robotics Research*, 34(7):1064–1082, 2015.
- X. Zhu, Z. Ghahramani, and J. Lafferty. Time-sensitive Dirichlet process mixture models. Technical report, Carnegie Mellon University, 2005.